The Future of IntegrOmics

Kristel Van Steen, PhD[2]

kristel.vansteen@ulg.ac.be

Systems and Modeling Unit, Montefiore Institute, University of Liège, Belgium

Bioinformatics and Modeling, GIGA-R, University of Liège, Belgium

Université de Liège

OUTLINE

- Do we have a motive and opportunity for data integration?
- What are the key components of data integration
- Do we also have the means?

Backbone references

- Hamid et al 2009. “Data integration in genetics and genomics: methods and challenges” (review article – appeared in *Human Genomics and Proteomics*)
- Davies et al 2009. “Integrative genomics and functional explanation” (presentation)

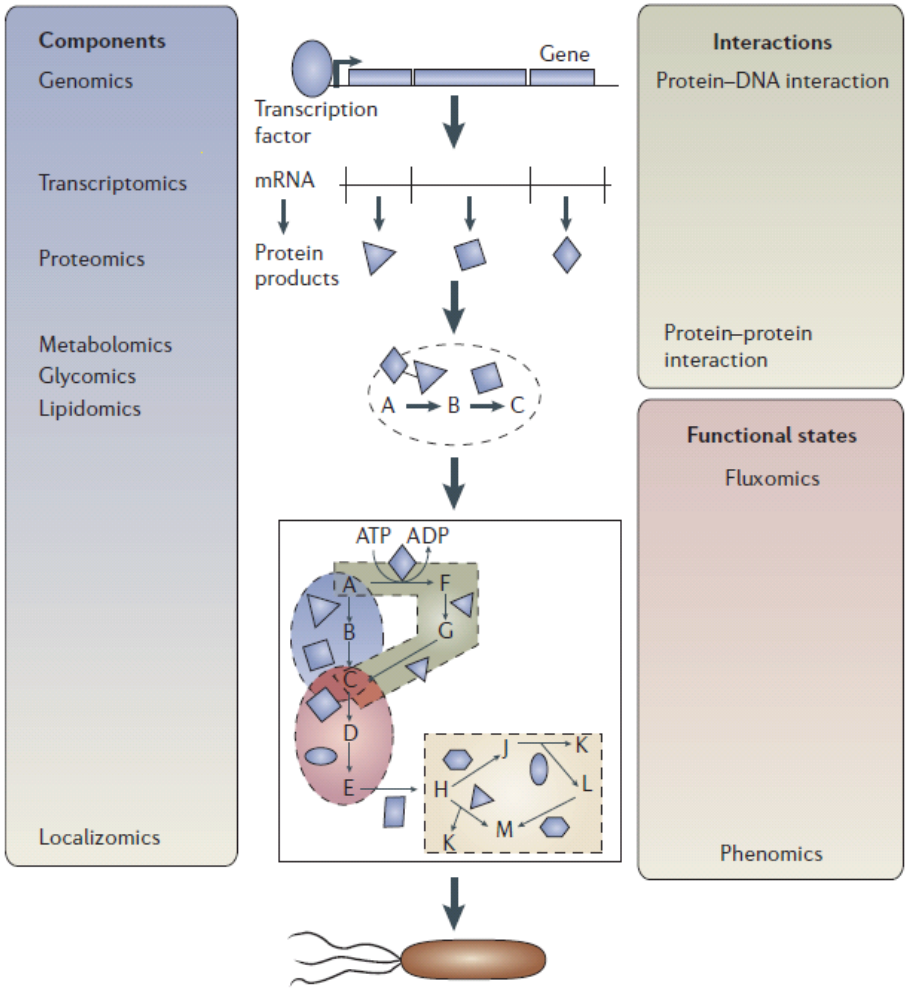
Motive and opportunity for data integration

Availability of different data resources

- Increased storage capacities and joint efforts to establish data banks with easy data access make available huge amounts of clinical, environmental, demographic data
- Rapid technological advances lead to various types of -omics data:
  - Genome (G)
  - Epigenome (E)
  - Transcriptome (T)
  - Proteome (P)
  - Metabolome (M)
  - Phenome (F)
  - Lipidome, glycome, interactome, spliceome, etc...

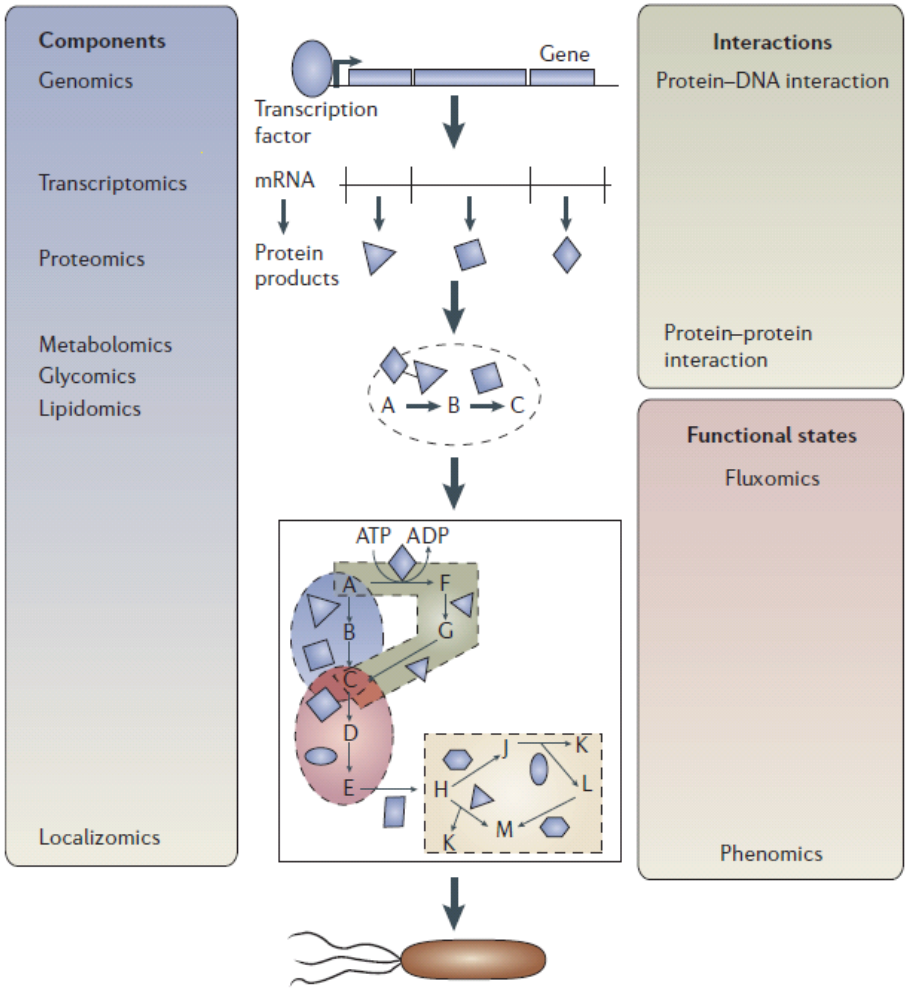
Omics: what's in a name?

- Omics data provide comprehensive descriptions of nearly all components and interactions within the cell.
- Three data categories:
  - Components
  - Interactions
  - Functional-states



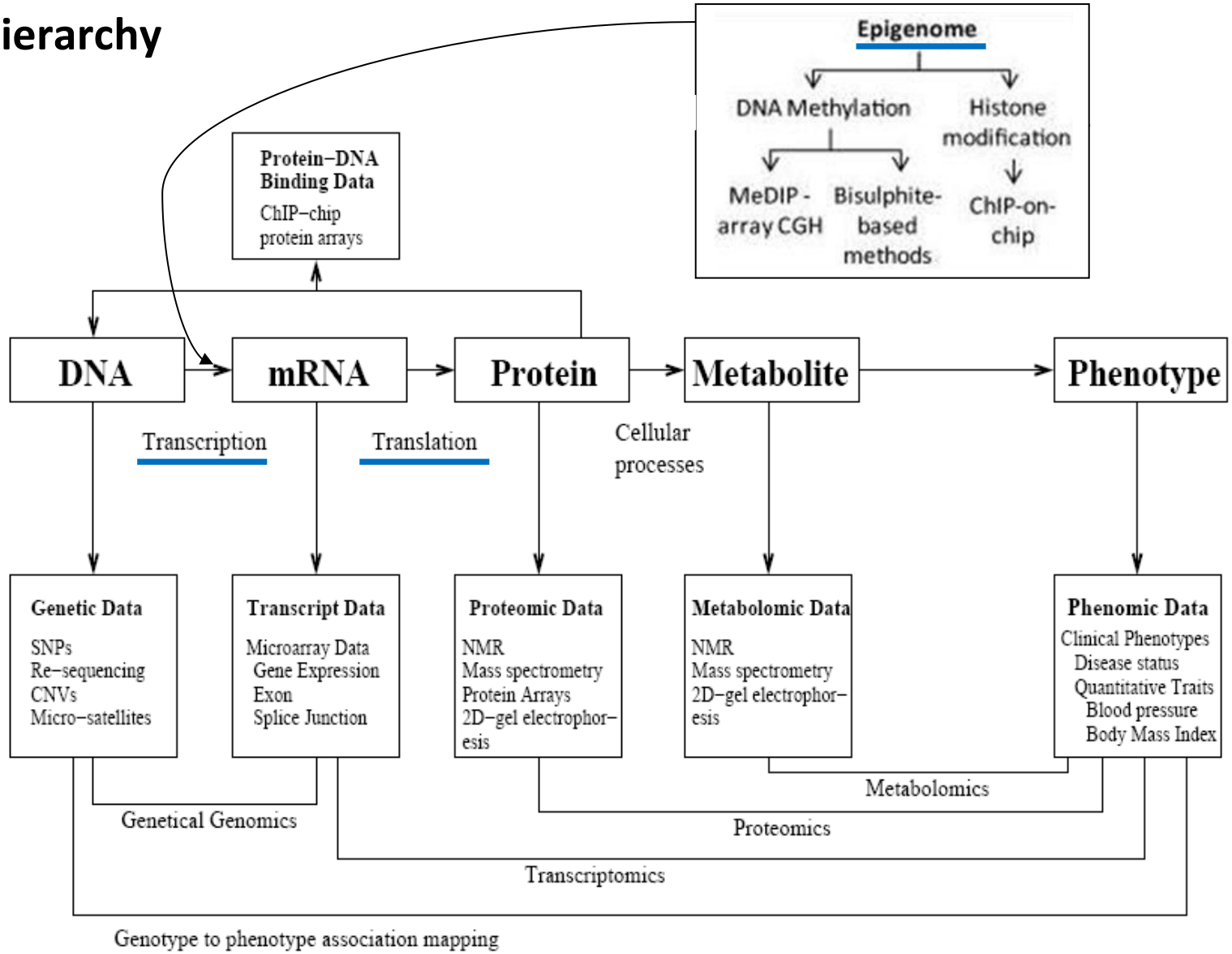
(Joyce and Palsson 2006)

Genomics: defined here as the study of the whole genome sequence and the information contained therein
Transcriptomics: provides information about both the presence and the relative abundance of RNA transcripts, thereby indicating the active components within the cell
Proteomics: aims to identify and quantify the cellular levels of each protein that is encoded by the genome
…



(Joyce and Palsson 2006)

Data hierarchy



(adapted from: Davies et al 2009, Integrative genomics and functional explanation)

NEWS IN FOCUS

PROSPECTS

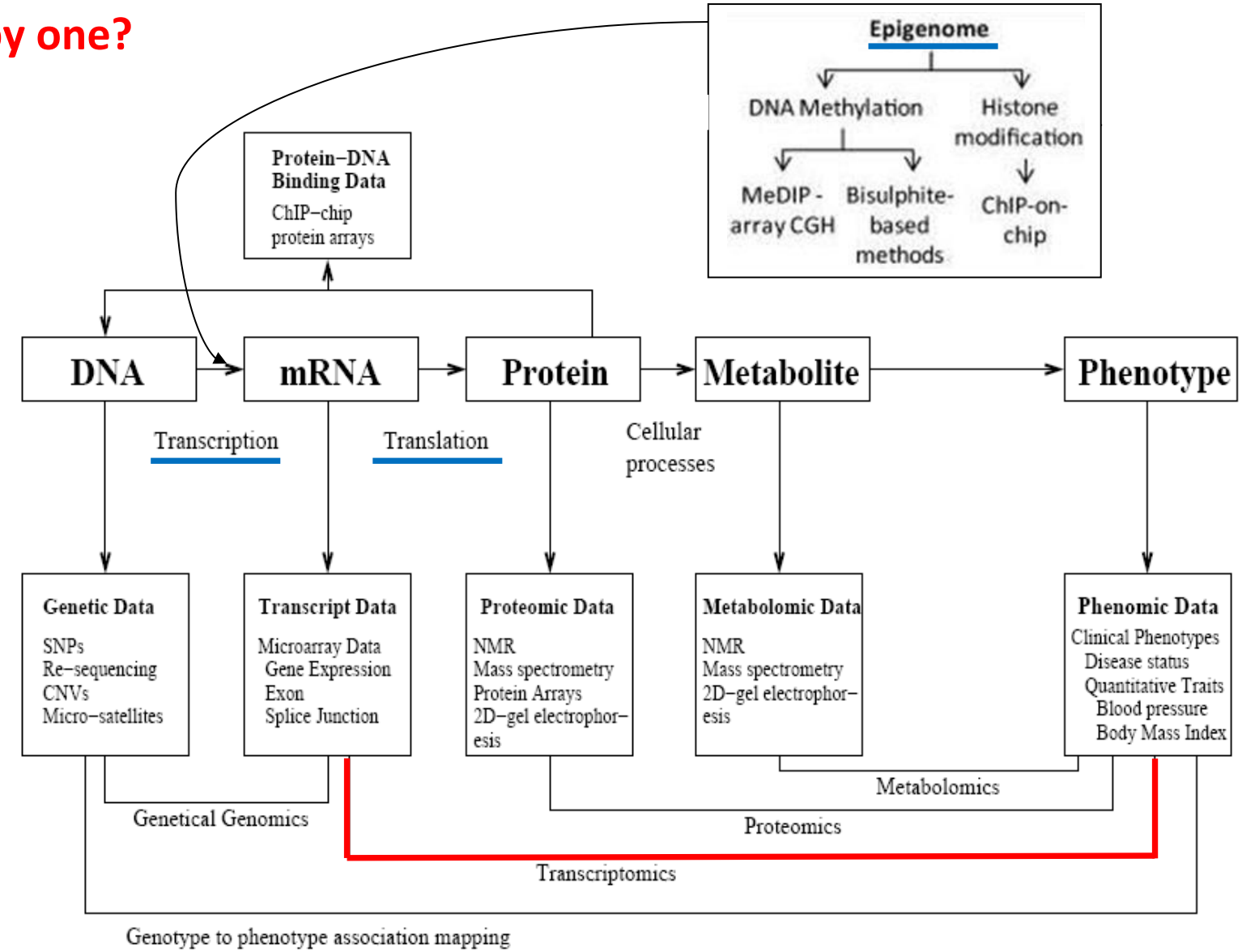
New year, new science

Nature *looks at key findings and events that could emerge from the research world in 2011.*

GWAS PROVE THEIR WORTH

Genome-wide association studies (GWAS) have uncovered plenty of links between diseases and particular regions of the genome, but frustratingly haven't revealed much about the biochemistry behind these associations. In 2011, expect to see real mechanistic insights explaining how genes, and non-coding regions, affect the medical conditions they have been linked with. Metabolism, obesity and diabetes are among the hottest targets.

One by one?



(adapted from: Davies et al 2009, Integrative genomics and functional explanation)

Université de Liège

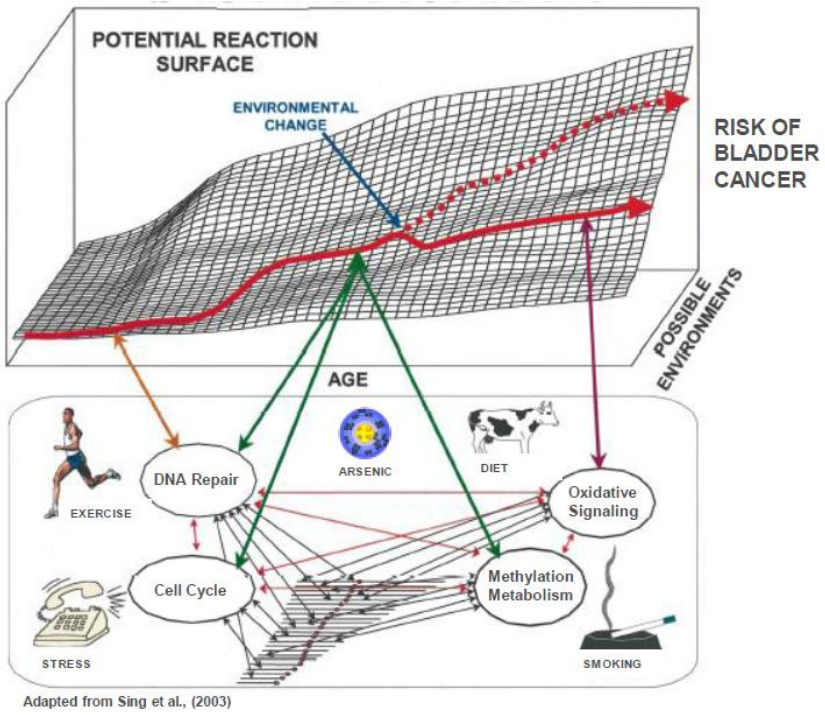
Functional genomics

- Full functional understanding of the etiology of a complex phenotype involves:
  - identifying the genetic, molecular, and environmental attributes that influence the phenotype, and
  - elucidating the biological pathway that fully defines the influence and describes how it occurs.
- The epigenome as the interface between the genome and the transcriptome, controlling long-term gene expression and integrating environmental signals, needs to be “integrated” into the functional genomics analysis.

A "forgotten" –ome data resource: the exposome

- The clinical management of cancer and other complex diseases can be substantially improved by "integrative" approaches.

.



(See also Wild 2005, Complementing the genome with an "exposome"…)

Motivation to integrate different data resources

- In general, distinct data types provide a different, only partly independent and complementary view on most complex disease related questions to date.
- The “genome” on its own “has turned out to be a relatively poor source of explanation for the differences between cells or between people” (Bains 2001).

So we have the **motive,** and the **opportunity** …

(Boston Globe)

… but do we also have the **means** for “integration” ?

(Mission Impossible @ google)

… but do we also have the **means** for “integration” ?

BMC Bioinformatics

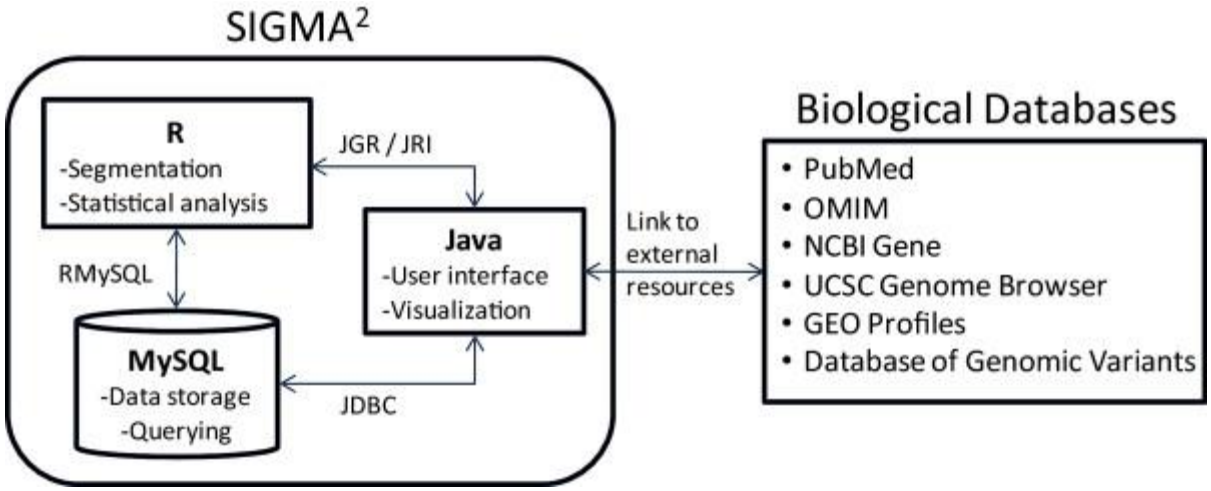
BioMed Central

Software

Open Access

SIGMA²: A system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes, and transcriptomes

Raj Chari*[1], Bradley P Coe[1], Craig Wedseltoft[1], Marie Benetti[1], Ian M Wilson[1], Emily A Vucic[1], Calum MacAulay[2], Raymond T Ng[3] and Wan L Lam[1]



BMC Bioinformatics

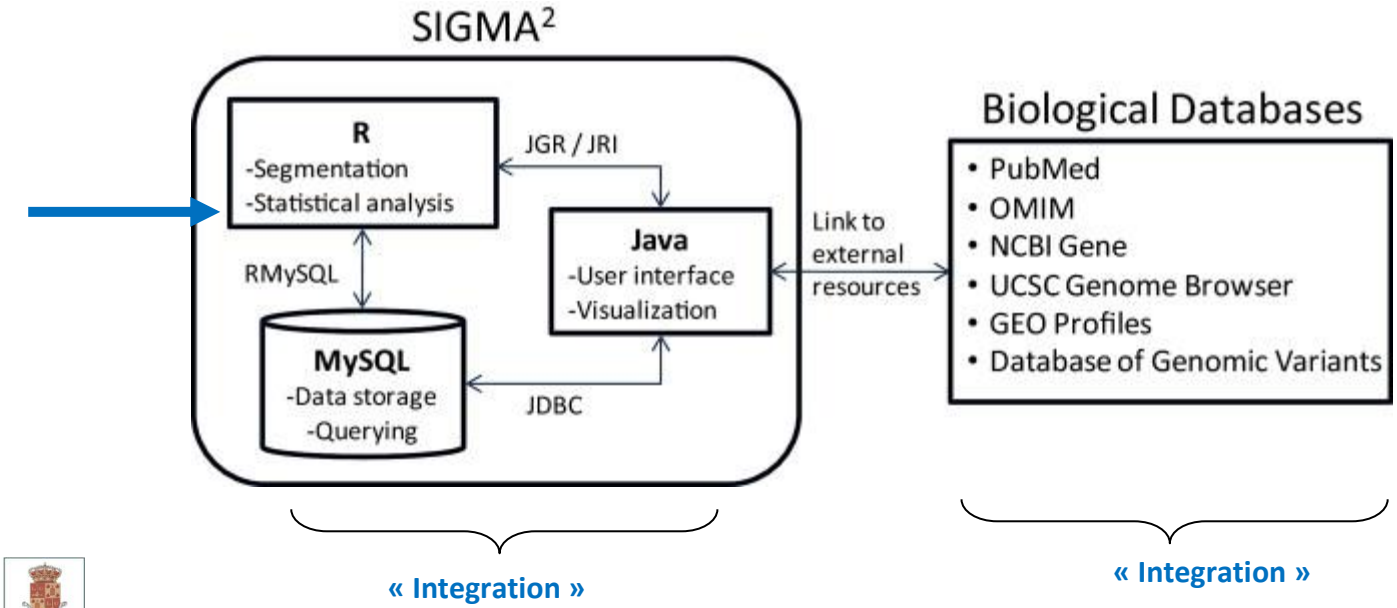
BioMed Central

Software

Open Access

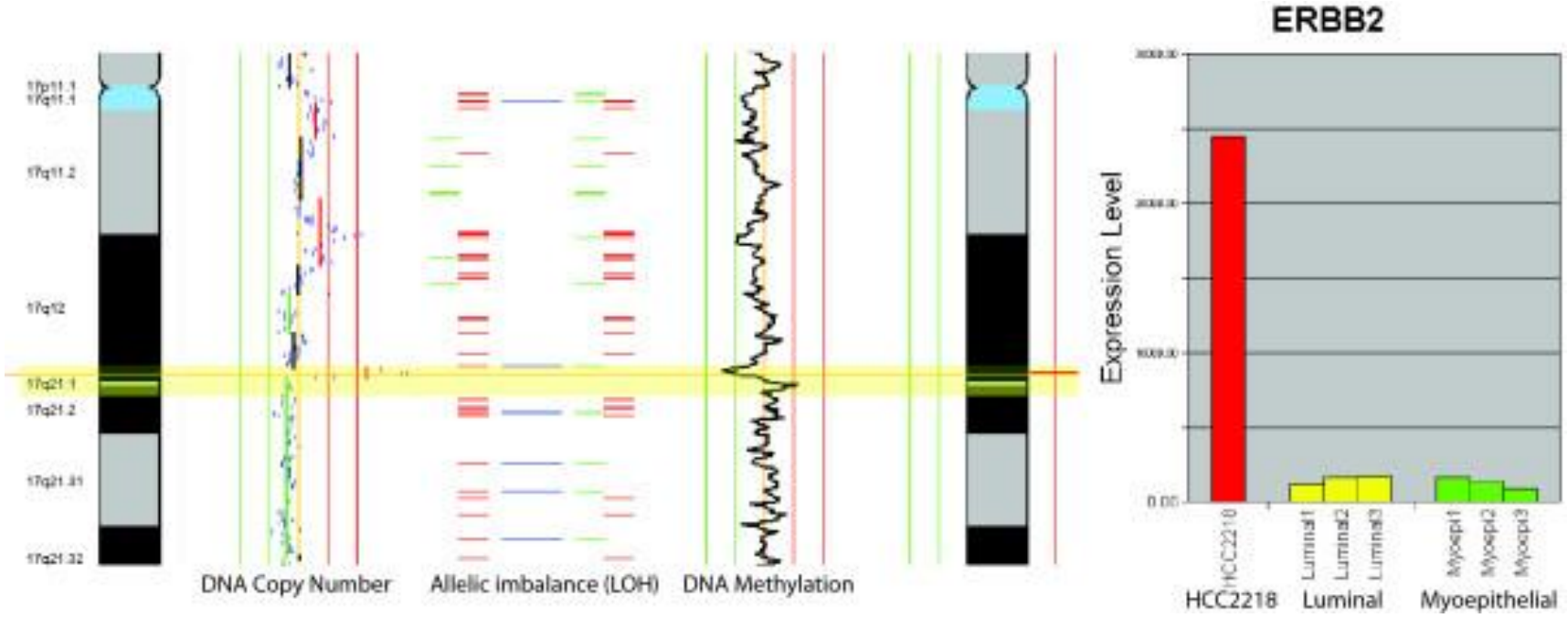
SIGMA²: A system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes, and transcriptomes

Raj Chari*[1], Bradley P Coe[1], Craig Wedseltoft[1], Marie Benetti[1], Ian M Wilson[1], Emily A Vucic[1], Calum MacAulay[2], Raymond T Ng[3] and Wan L Lam[1]



Integration as "obtaining a multi-dimensional perspective" ?

of chromosome 17 of the HCC2218 breast cancer cell line



Copy number, LOH, and DNA methylation, and profiling identifies an amplification of ERBB2 coinciding with allelic imbalance and loss of methylation. The expression of HCC2218 is significantly higher than a panel of normal luminal and myoepithelial cell lines.

(Breast Cancer Res. 2006;8(5):R56.)

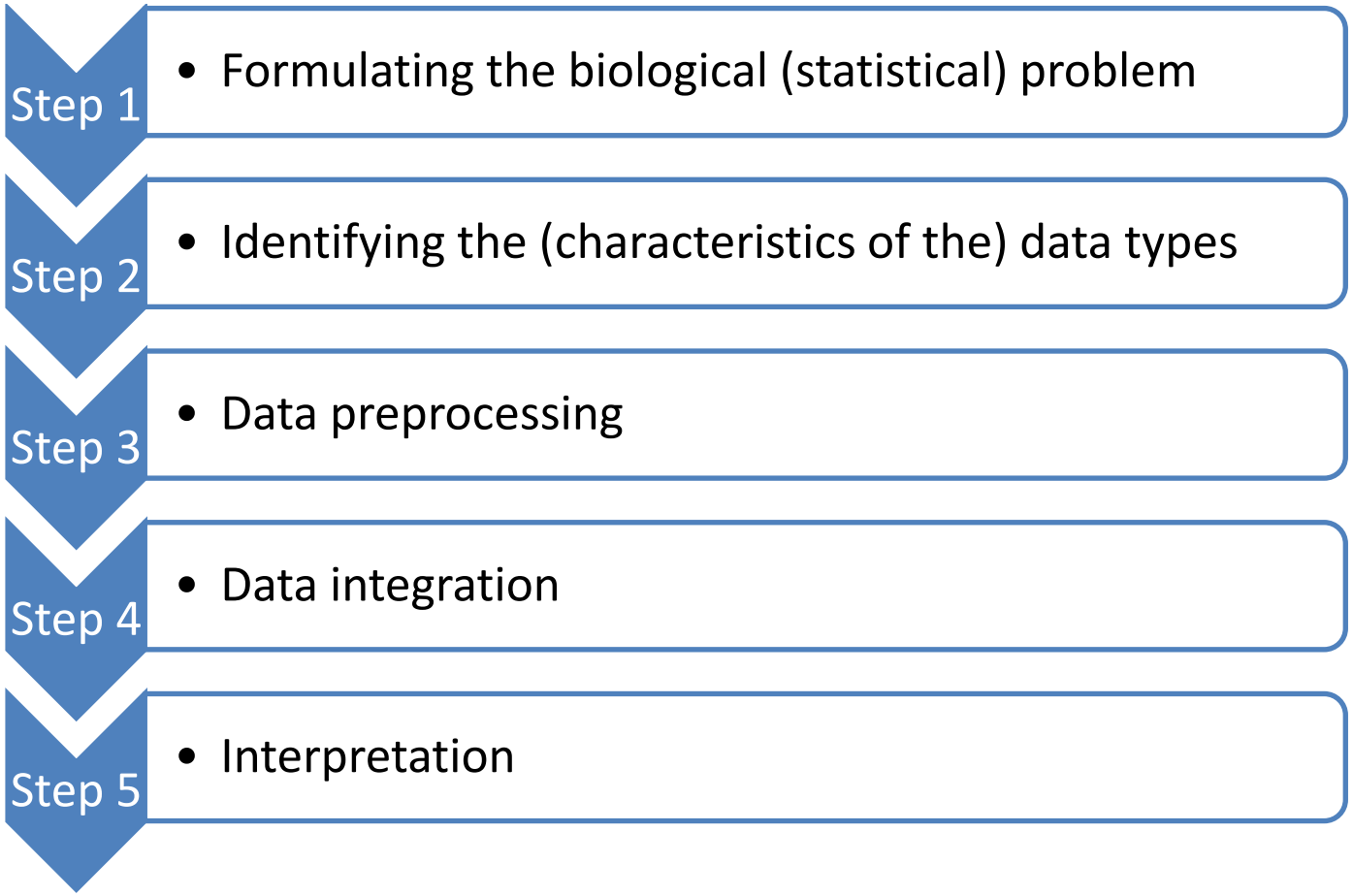
Key components of data integration

Data integration: what's in a name?

Definition

- **Narrow:** "Process of statistically combining data from different sources to provide a unified view of the whole genome and make large-scale statistical inference" (Lu et al 2005).
- **Broad:** "Combining evidences from different data resources, as well as data fusion with biological domain knowledge, using a variety of statistical, bioinformatics and computational tools" (Van Steen)

Key components of (statistical) data integration



Step 1

- Formulating the biological (statistical) problem

- Traditional biological research questions are for the most part hypothesis-driven where one performs experiments to answer specific biological hypotheses.
- In modern genomics, it is increasingly accepted to generate data in a relatively hypothesis-free setting where different questions can be postulated on the pool of data and data are mined with a variety of computational and statistical tools with the hope of discovering new knowledge.

Examples

Genomics	Transcriptomics	Proteomics	Metabolomics	Protein–DNA Interactions	Protein–protein Interactions	Fluxomics	Phenomics
Genomics (sequence annotation)	• ORF validation • Regulatory element identification[74]	• SNP effect on protein activity or abundance	• Enzyme annotation	• Binding-site identification[75]	• Functional annotation[79]	• Functional annotation	• Functional annotation[75,103] • Biomarkers[115]
	Transcriptomics (microarray, SAGE)	• Protein: transcript correlation[10]	• Enzyme annotation[108]	• Gene-regulatory networks[76]	• Functional annotation[80] • Protein complex identification[82]		• Functional annotation[102]
		Proteomics (abundance, post-translational modification)	• Enzyme annotation[99]	• Regulatory complex identification	• Differential complex formation	• Enzyme capacity	• Functional annotation
			Metabolomics (metabolite abundance)	• Metabolic-transcriptional response		• Metabolic pathway bottlenecks	• Metabolic flexibility • Metabolic engineering[109]
				Protein–DNA interactions (ChIP–chip)	• Signalling cascades[86,102]		• Dynamic network responses[84]
					Protein–protein interactions (yeast 2H, coAP–MS)		• Pathway identification activity[88]
						Fluxomics (isotopic tracing)	• Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)

Addressing biological questions at the systems level

(Joyce and Palsson 2006)

Step 2

- Identifying the (characteristics of the) data types

- Current data integration methods fall into two different categories:
  - integrating similar data types (across studies) or
  - integrating heterogeneous data types (across studies as well as within studies).
- Data are classified as heterogeneous if two or more fundamentally different data sources are involved.
  - E.g., the problem to develop a predictive model based on different genomic data (SNP, gene expression, protein, sequence) as well as clinical data, clearly involves heterogeneous data sources.

Step 2

- Identifying the (characteristics of the) data types

- Data characterization refers to finding preliminary evidences for
  - layers of information within and between data sets
  - noise patterns (e.g., depending on the technology, the platform, the lab, and many other systematic and random errors)
  - precision
  - accuracy (e.g., error proness)
  - reliability
  - dimensionality
  - quality
  - intrinsic properties (e.g., small sample sizes, standard formats)

Step 2

- Identifying the (characteristics of the) data types

- Whether data are of similar or heterogeneous type, the issue of quality (each data sources is unavoidably subject to different levels of noise) and informativity is of great importance.
- Therefore, the concept of weighting the data sources with quality and/or informativity scores becomes an essential component of the framework.
- Step 2 to data integration is as important as a classical Exploratory Data Analysis (EDA) in statistical inference practice.

Step 3

- Data preprocessing

- Genomic data are subject to different noises and errors, and a number of critical steps are required to preprocess raw measurements.
- Approaches for preprocessing vary depending on the type and nature of data (e.g., arrays: background correction, normalization, quality assessment, which may differ from one platform to another)
- Data preprocessing can be done at any step of the data integration process, most often at the initial stage (e.g., see before: data prepping, but also data matching /data imputation) and prior to statistical analysis (e.g., checking whether the assumptions of the statistical method are met and taking actions accordingly).

Step 4

- Data integration

- Data from different sources can be integrated at three different stages, and the choice for either of these depends on the biological question, the nature, and type of data as well as the availability of original data.

- **Early stage:** Merging data from different studies or experiments to increase sample size. Merging weighted (e.g., including sample size, quality and/or informativity scores) data.
  - Note that attaching weights to the data does not change the general format and nature of the resulting data.

Step 4

- Data integration

- **Intermediate stage:** Transforming individual data sources into another (common) format and/or dimension before combining them.
  - For example, in class identification problems (subphenotyping), one might convert the data into similarity matrices such as the covariance or correlation matrix and combine these similarity matrices for better "clustering".

- **Final stage:** Final statistical results from different studies are combined.
  - This stage includes, among others, meta-analytic techniques where one typically combines effect sizes or p-values across studies.

Which classes are often "integrated" in analysis?

G = Genome, E = Epigenome, T = Transcriptome, P = Proteome, M = Metabolome, F = Phenome

- ***Analysis of single sources of data***
  - Species Level Genomic Variation Data; (G)
  - Human Genetic Variation Data; (G)
  - Molecular Quantities; (T), (M), (P)
- ***Analysis of phenotype with another source of data***
  - Analysis of phenotype with genetic data; **(G+F)**
  - Analysis of phenotype with molecular data; **(F + T)**, (F + P), (F + M)
  - Analysis of genetic data with molecular data; **(G + T),** (G + P), (G + M)

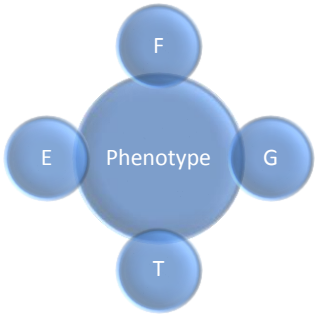
Which classes are often "combined" in analysis?

G = Genome, E = Epigenome, T = Transcriptome, P = Proteome, M = Metabolome, F = Phenome

- ***Analysis with multiple molecular data types***
  - (T + P), (T + M), (M + P)

- ***Analysis of all data types across multiple species***

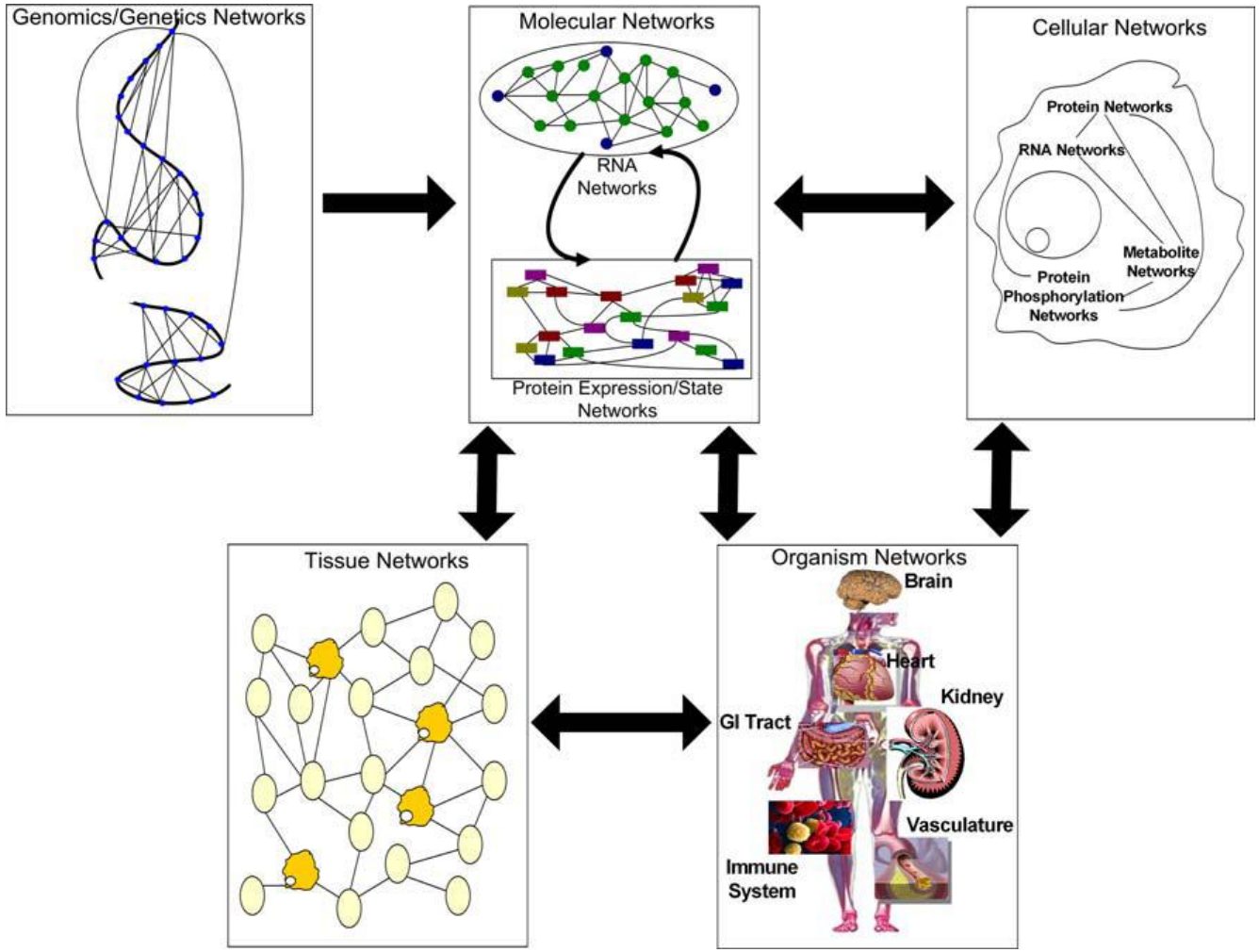
- ***Integrated analysis of phenotype (F) with at least two other sources of data***



How?

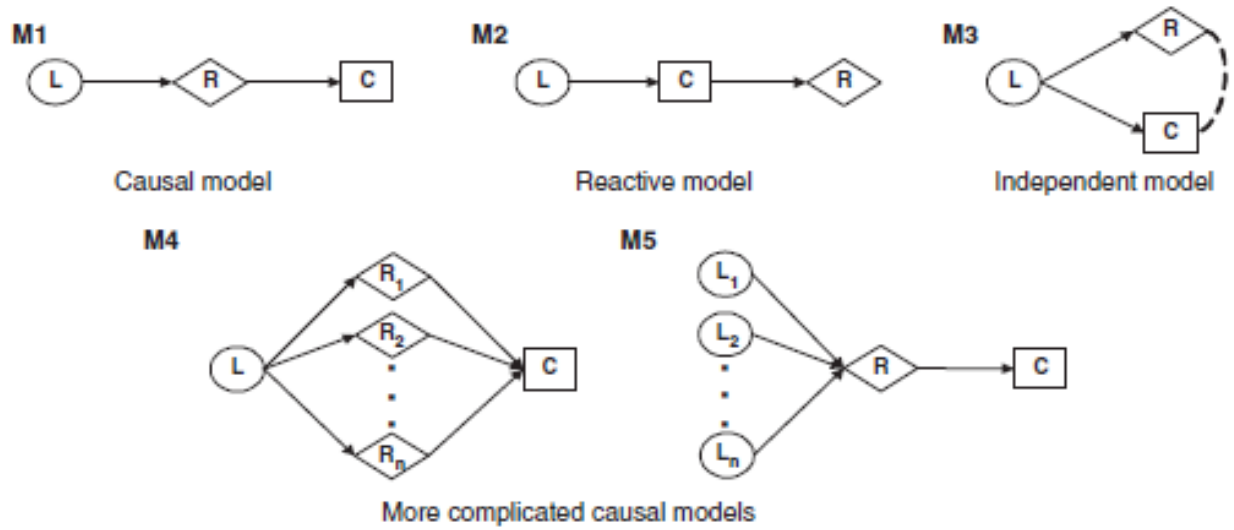
- Analyze phenotype separately and then compare vs truly multivariate analysis of phenotypes
- Integrated network analysis (e.g., combining statistical interaction networks)

Flow of information in biological systems through a hierarchy of networks



(Sieberts et al 2007)

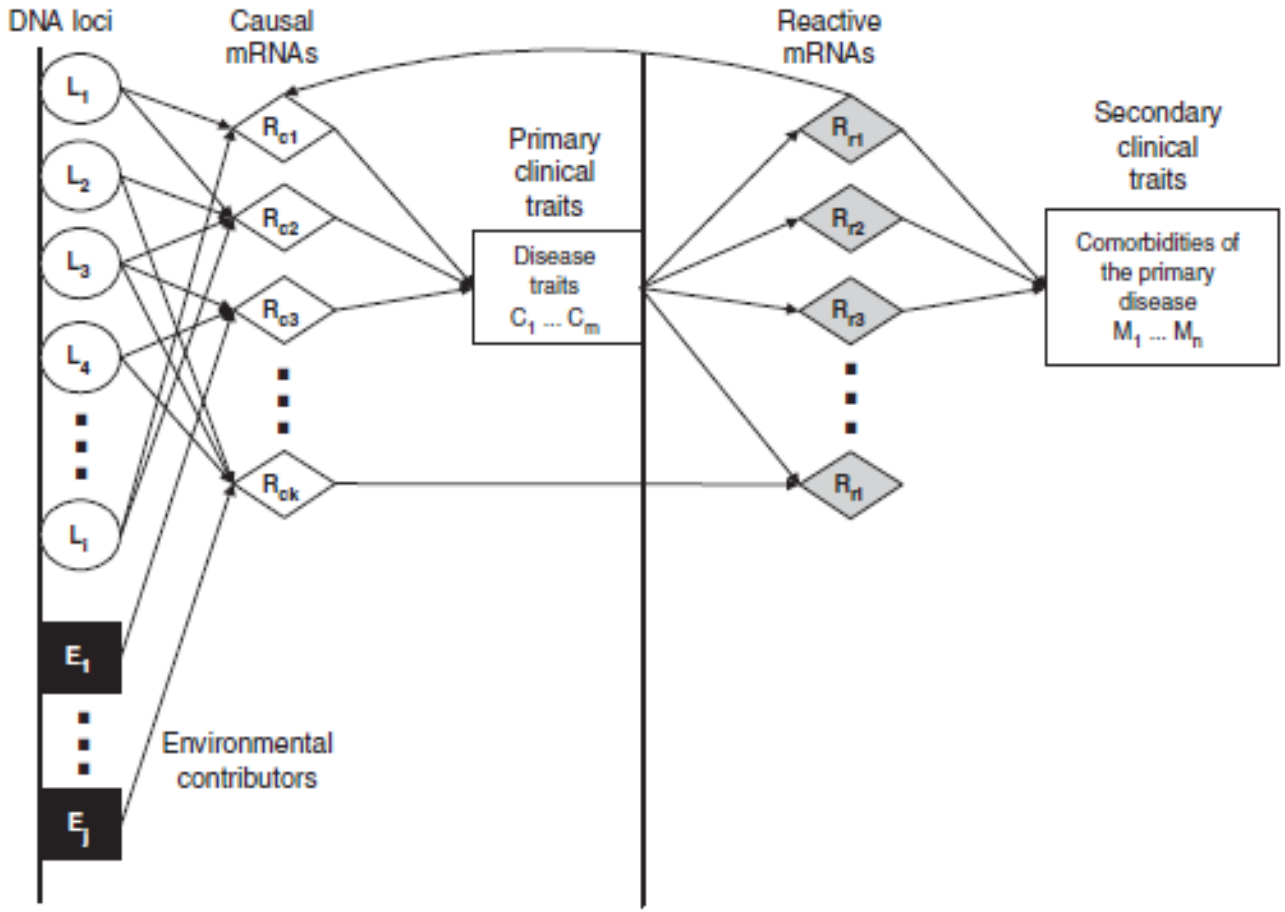
QTL data to infer relationships between RNA levels and complex traits



(Schadt et al 2005)

- Several graphical models to represent possible relationships between QTLs, RNA levels and complex traits once the expression of a gene (R) and a complex trait (C) have been shown to be under the control of a common QTL (L).

Hypothetical gene network for disease traits and related comorbidities



(Schadt et al 2005)

Step 5

- Interpretation

- Is about "understanding" the problem that was initially posed.
- Involves post-linking to several external biological data bases
- Non-standard (!!!) to do's when "integrating" evidences from these biological data bases (e.g., finding evidences of "interactions")
  - Assess and incorporate "optimal" scoring systems to accumulate evidence from these data bases
  - Allow for uncertainty involved in the data source entries
  - Acknowledge the complementary characteristics of each of the available data sources
  - Allow for different assignment strategies (e.g., from genetic variants to genes)

Step 5

- Interpretation

- Interpretation often involves functional explanation (as part of functional genomics)
- Experimental validation in theory remains necessary, but in practice often is infeasible!
- Visual analytics: There is a huge challenge in visualizing the steps of and the results from an integrated analysis

Do we have the means?

General methodologies

- Before discussing main stream type of developments, recall that – whatever the strategy - biological knowledge (when available and reliable) should be used to support the integration process at every stage possible (cfr., early, intermediate, late)
- The 3 stages above inherently give pointers towards an analytic solution.
- Naively:
  - Early, final stage:
    - POOL the data
    - Key word = WEIGHTS
  - Intermediate stage:
    - REDUCE the data / select representatives
    - Key word = PATTERN DISCOVERY

Pattern discovery methods « within and across »

"Combining" heterogeneous data sources should not be done carelessless:

Questions	**Examples of related concept**
How "informative" is each data set?	Entropy
What are the most "important" features in each set?	Prediction, information content
Are these modified by internal or external factors?	Interactions, networks
What is the amount of "overlap" between different data sets?	Synergy, redundancy

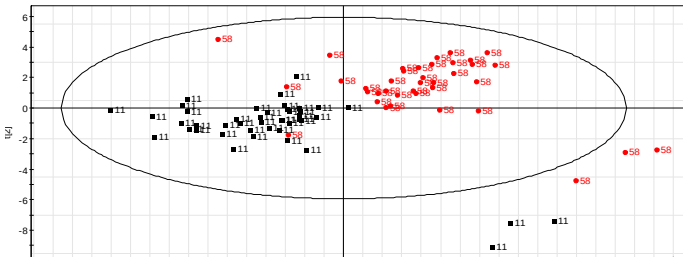
- In general, pattern discovery techniques usually have in common that they are “unsupervised” in nature, referring to the property that “informative reduction” of the data is obtained solely from the data itself and not from a priori knowledge or any classification scheme.
- Classical approaches can be grouped as dimension-reduction techniques (e.g., components-based: principal components) and data mining techniques (e.g., clustering, (conditional) tree/forest/jungle-building).
- Less classical approaches involve network-based strategies
- Challenges:
  - (Differential) sparseness + noise + scales + dimensionality + …
  - Combining / integrating over data types

Multivariate dimension-reduction techniques

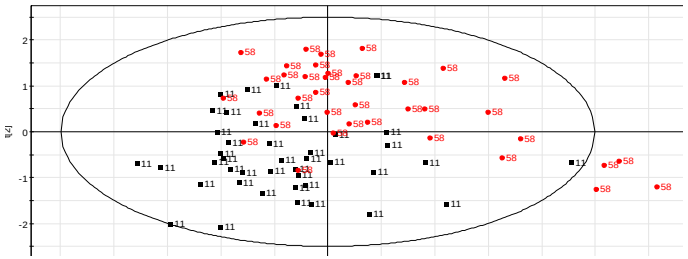
- Those methods are used depending on the type of data at hand whether variables are quantitative (numerous) or qualitative (categorical or nominal):
  - **Principal component analysis** (PCA) when individuals are described by quantitative variables;
  - **Correspondence analysis** (CA) when individuals are described by two categorical variables that leads to a contingency table;
  - **Multiple correspondence analysis** (MCA) when individuals are described by categorical variables.

Multivariate dimension-reduction techniques

- Simply combining “components” from different data types?

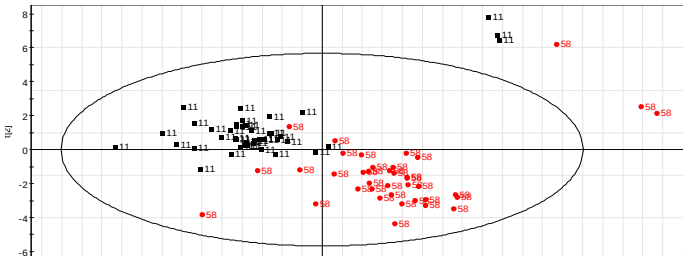


Non-omics (20)

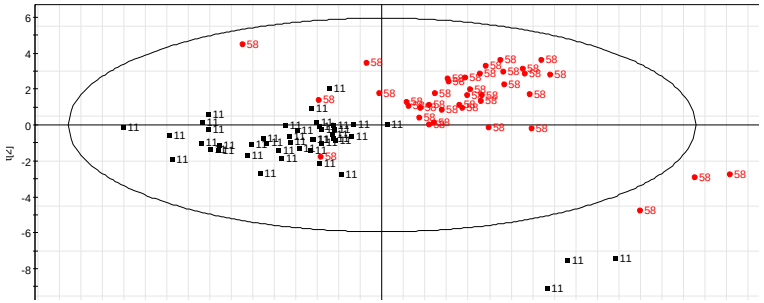


Transcriptomics (20 PCs)

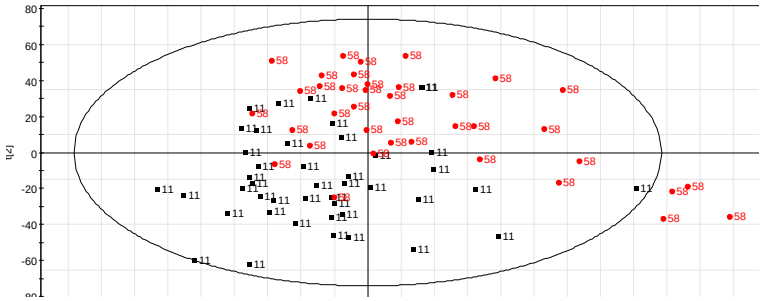
Like a mirror image!!!



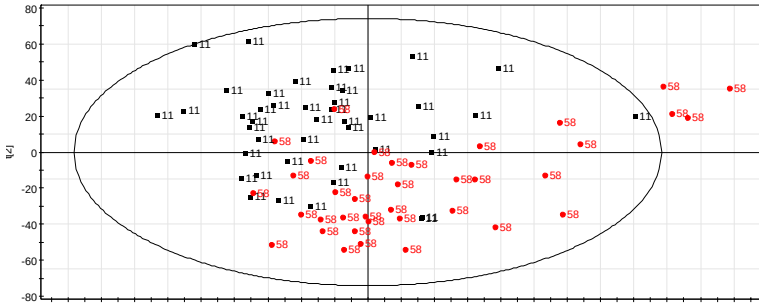
Combined (40)



Non-omics (20)



Transcriptomics (12,488)



Combined (12,508)

(plots: Lee et al; NISS Metabolomics Workshop, 2005)

Multivariate dimension-reduction techniques

- Not frequently used in omics - allowing more structure on the data
  - Groups of variables:
    - E.g., **Multiple factor analysis** (MFA) - (Escofier and Pagès 1988)
      - MFA allows to analyse several groups of variables which can be quantitative and/or categorical (e.g., multiple phenotypes, genotypes, transcripts)
      - The heart of MFA is a PCA in which group-weights are assigned to the variables
  - Hierarchy on the variables:
    - E.g., **Hierarchical multiple factor analysis** (HMFA) – (LeDien and Pagès 2003)
      - E.g., 2 groups of variables (phenomics and other omics), where the omics group is subdivided in transcriptomics, genomics, etc.
  - Groups of individuals
    - E.g., **Dual multiple factor analysis** (DMFA) – (Lê and Pagès 2007)

Multivariate networks-based techniques

- Networks and deriving basic network properties are well established in a lot of omics fields separately
- The key is how to define a connection/bond between nodes:
  - via statistical "association" (e.g., statistical epistasis networks)
  - via general "association"
- When multiple omics fields are considered jointly, there are two options:
  - Combining networks, accounting for different degrees of granularity, informativeness and precision,
  - Compairing networks (descriptive versus hypothesis testing)

Multivariate networks-based techniques

OPEN ACCESS Freely available online PLoS one

Network Properties of Complex Human Disease Genes Identified through Genome-Wide Association Studies

Fredrik Barrenas[1☯*], Sreenivas Chavali[1☯], Petter Holme[2,3], Reza Mobini[1], Mikael Benson[1]

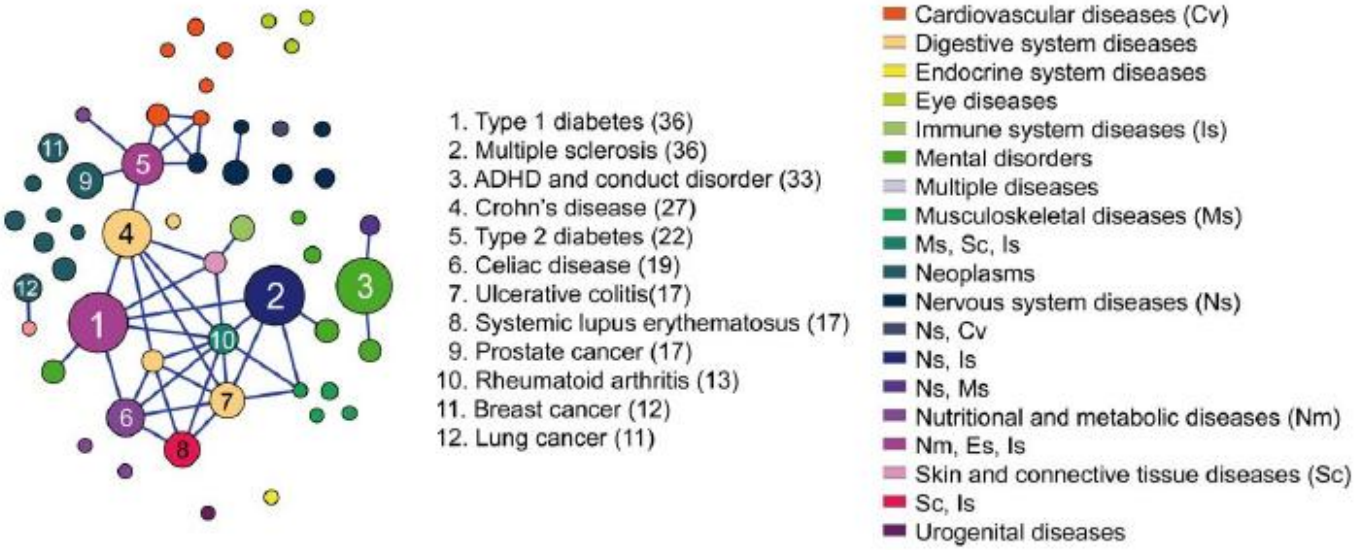
1 The Unit for Clinical Systems Biology, University of Gothenburg, Gothenburg, Sweden, **2** Department of Physics, Umeå University, Umeå, Sweden, **3** Department of Energy Science, Sungkyunkwan University, Suwon, Korea

Abstract

Background: Previous studies of network properties of human disease genes have mainly focused on monogenic diseases or cancers and have suffered from discovery bias. Here we investigated the network properties of complex disease genes identified by genome-wide association studies (GWAs), thereby eliminating discovery bias.

Principal findings: We derived a network of complex diseases (n = 54) and complex disease genes (n = 349) to explore the shared genetic architecture of complex diseases. We evaluated the centrality measures of complex disease genes in comparison with essential and monogenic disease genes in the human interactome. The complex disease network showed that diseases belonging to the same disease class do not always share common disease genes. A possible explanation could be that the variants with higher minor allele frequency and larger effect size identified using GWAs constitute disjoint parts of the allelic spectra of similar complex diseases. The complex disease gene network showed high modularity with the size of the largest component being smaller than expected from a randomized null-model. This is consistent with limited sharing

- Example 1: Complex Disease Network (CDN)



(Barrenas et al 2009)

Complex Disease Network (CDN). Each node is a complex disease studied in GWAs with the link representing sharing of disease genes. The color of the nodes corresponds to disease class as identified using MeSH (Medical Subject Headings) terms as given on the right side. Notably, complex diseases are hard to define using single MeSH term. The node size refers to the number of associated genes identified. Diseases with most number of associated genes identified through GWAs are listed on the right side with the numbers in the parenthesis indicating the number of associated genes.

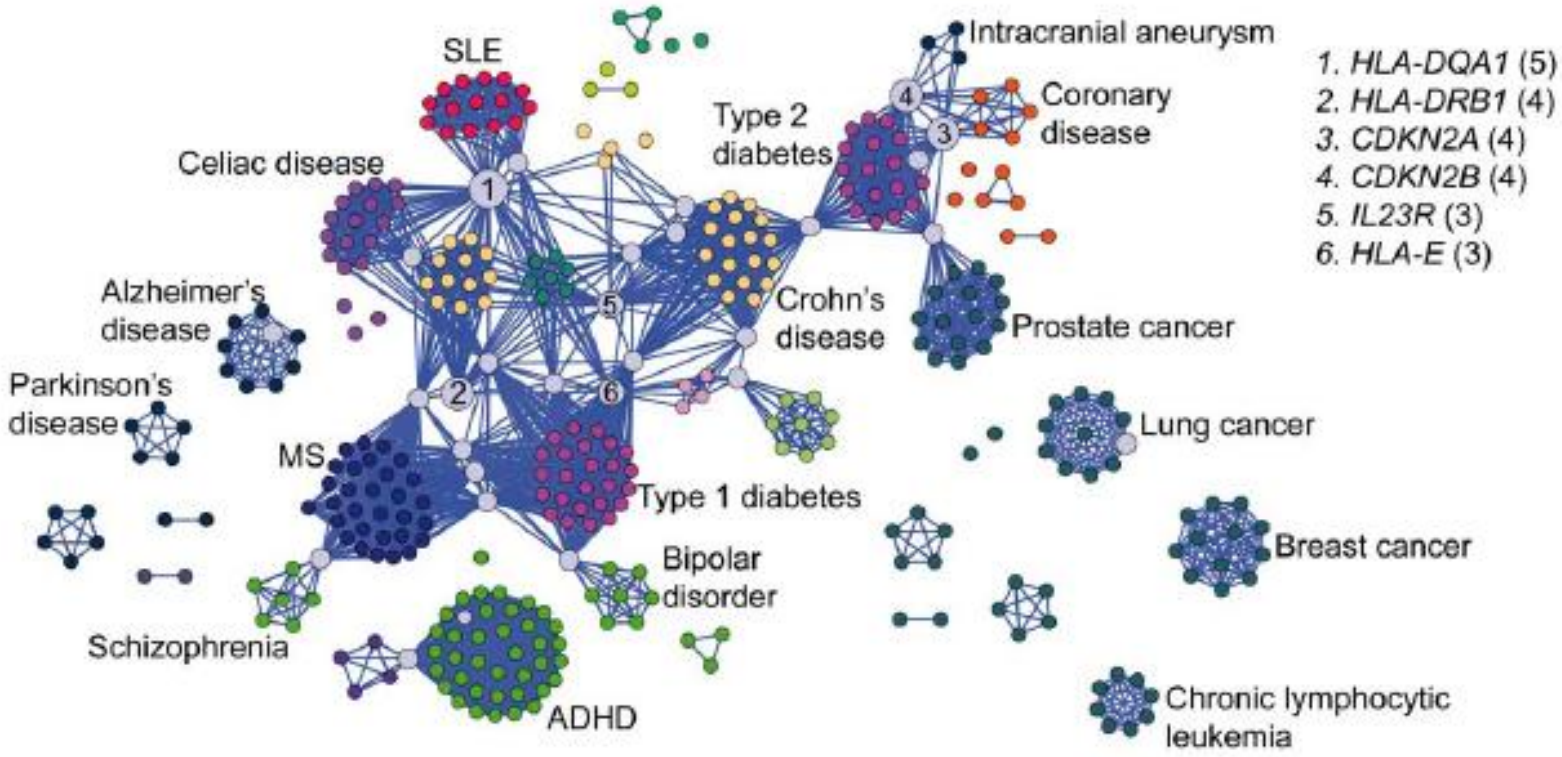
Pleiotropy

(= occurs when one gene influences multiple phenotypic traits)

Phenocopy

(= "environm. induced", nonhereditary variation in an organism, closely resembling a genetically determined trait)

- Example 2: Complex Disease Gene Network (CGN)



Complex Disease Gene Network (CGN): Each node represents a gene and connections between two genes represent their association with the same disease. The node size refers to the number of diseases a gene is associated with. Genes associated with many diseases are listed on the top right side with the number of diseases they are implicated with, in the parenthesis. A node (highlighted in gray) each in lung cancer and Alzheimer's disease gene cluster are singular associations in idiopathic pulmonary fibrosis and narcolepsy respectively.

(Barrenas et al 2009)

In Conclusion

1996: Microsatellites & linkage

development of NIDDM in susceptible individuals. These studies will lead to a better understanding of the causes of NIDDM and its complications, and improved diagnosis and treatment of this chronic disorder.

A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2

2007: SNPs & genome-wide association

ing a more complex scenario of pleiotropic effects. We anticipate that identification of the causal variants at these genetic loci and their functional consequences will reveal unexpected players in T2DM pathogenesis, and will point to novel mechanisms and targeted therapeutics.

A genome-wide association study identifies novel risk loci for type 2 diabetes

Vol 445 | 22 February 2007 | doi:10.1038/nature05616

2010: Rare variants & sequencing

"We are moving genomic research into a new phase," said Mark McCarthy, Professor of Diabetes at the University of Oxford, who is leading the British side of the type 2 study. "We will be able to pin down many more genetic effects on disease, to understand the biology more completely and give the pharmaceutical industry new targets for drugs."

January 5, 2010

Landmark DNA study of 3,000 people to unlock mystery of type 2 diabetes

TIMESONLINE

(slide : Lon Cardon)

Université de Liège

Take-home message

- To date, the global genome-wide studies describe systems of a size that cannot be modeled to the detailed level of biological systems
- Functional interpretation is attempted by integrative studies and systems biology (yet, both are still too high level to provide full functional explanations at a molecular or atomic level!)
- Along with the continuing rise of high throughput studies, there will be a niche for improved mathematical modeling and advances on integrative biology (top-down) as well as systems biology (bottom-up) and its relations

(Davies et al 2009, Integrative genomics and functional explanation)

Food for thought

- Use biological info at all levels?
  - Not so much of the biology is known …
- Validation lab experiment may not be true in practice
  - Experiments involving IntegrOmics are highly complex …
- How to bring in dynamics:
  - Fuzzy networks: nodes are rule bases … / Association Rule Learning
- Will IntegrOmics facilitate development of better prediction models?
  - Predict whether the disease will occur
  - Predict effectiveness of treatment of hazardous event (e.g., flu)
  - Predict the development of a disease, conditional on having it.